A probabilistic interpretation of the constant gain algorithm

Michele Berardi

University of Manchester

19 May 2019

A probabilistic interpretation of the constant gain algorithm

Michele Berardi
The University of Manchester

May 19, 2019

Abstract

This paper proposes a novel interpretation of the constant gain learning algorithm through a probabilistic setting with Bayesian updating. Such framework allows to understand the gain coefficient in terms of the probability of changes in the estimated quantity.

Key words: Bayesian learning, adaptive learning, constant gain.
JEL classification: C63, D83, D84, D90, E70.

A probabilistic interpretation of the constant gain algorithm

1 Introduction

Adaptive learning has been used extensively in the macroeconomic literature to depart from the assumption of rationality. An extensive treatise of adaptive learning in macroeconomics is provided by Evans and Honkapohja (2001). In this framework, learning algorithms are used to model how agents recursively estimate relationships between variables of their interest, i.e., parameters in economic models. Two main forms of the learning algorithm are used in the literature: decreasing gain (DG) and constant gain (CG) learning.

The most common instance of a DG algorithm is the recursive least squares, where the gain coefficient is set equal to $1/t$, t being the time period of the estimate, and all observations are thus weighted equally: this is suitable for estimating quantities that are believed to be constant over time. With a constant gain algorithm, instead, more recent observations receive a higher weight, and the weights decrease geometrically with time: this is usually employed when the estimated parameters are believed to change over time, as it allows for better tracking.

A growing literature in applied macroeconomics has used CG learning to explain a range of features, from the rise and fall of U.S. inflation in the 70s and 80s (in particular, the seminal works of Sargent (1999) and Sargent at el. (2006)) to the causes of business cycles (e.g., Milani (2011) and Eusepi and Preston (2011)). Though there is no direct evidence of the appropriate value for the gain parameter, Berardi and Galimberti (2017) provide a thorough discussion of the role and estimate bands for the gain parameter in macroeconomic applications. In general, higher gains imply faster reaction to changes, but more volatile estimates.

The CG algorithm is a "reduced form" learning model, which could be derived as an optimal solution of inference in a number of underlying frameworks. For example, Muth (1960) has shown how adaptive expectations can be optimal under certain assumptions about the structure of the variable being forecasted. A CG algorithm for estimating the (mean) value of a variable, in fact, implements adaptive expectations, and as such it provides optimal forecasts under conditions specified in Muth (1960). Those conditions are quite restrictive on the underlying process for the variable being forecasted, which must be representable as an infinite sum of current and past exogenous disturbances, with appropriate weights related to the gain parameter.

A CG algorithm can also be obtained through a Kalman filter model, which implements Bayesian updating in a state-space framework, with appropriate initial conditions. It is well known that with a time-invariant state-space model, the Kalman gain converges to a constant: choosing such constant as initial value for the gain, the Kalman filter gives rise to a CG algorithm. The natural interpretation of such gain coefficient is usually in terms of the variances of disturbances in the measurement and transition equations.

I propose here instead an interpretation of the CG learning algorithm through a probabilistic setting where Bayesian learners estimate recursively the value of an unobservable variable through a signal. The underlying process for the variable being forecasted is not specified a priori through a parametric model, and only its probabilistic structure is defined. This framework allows for a novel interpretation of the gain coefficient in terms of the probability of changes in the estimated quantities. I then assess the values of various gain coefficients used in empirical studies against this background, deriving some implications on the underlaying frequency of changes in the estimated parameters.

2 Constant gain algorithm and Bayesian learning

Recursive learning algorithms can represent optimal learning behavior under certain assumptions about the underlying quantities to be learned. The simplest example is that of a constant, for which a decreasing gain algorithm with gain equal to $1/t$ is optimal, as it allows to estimate the sample mean. If the underlying variable to be estimated is instead time-varying, the literature suggests that a constant gain algorithm should be used, as it puts more weight on more recent observations and thus allows for better tracking.

2.1 Constant gain algorithm

Suppose agents need to estimate the (time-varying) mean of a random variable x_t over time. Denoting $\tilde{x}_t$ such estimate, the CG algorithm takes the form

$$\begin{aligned} \tilde{x}_t &= \tilde{x}_{t-1} + g\left(x_t - \tilde{x}_{t-1}\right) = \left(1-g\right)\tilde{x}_{t-1} + gx_t \qquad (1) \\ &= g\sum_{j=2}^{t}\left(1-g\right)^{t-j} x_j + \left(1-g\right)^{t-1} x_1, \qquad (2) \end{aligned}$$

where I have used the assumption $\tilde{x}_1 = x_1$, since no previous information is available to agents. The constant gain g determines the weight put on past observations, as

$$b_1^t = (1-g)^{t-1} \tag{3}$$
$$b_j^t = g(1-g)^{t-j} \tag{4}$$

for $j = 2, ..., t$, where b_j^t denotes the weight put at time t on time $j \leq t$ observation.

The same relationship between gains and weights holds also for a multivariate model where agents estimate a vector of time-varying coefficients through a linear regression model. See Berardi and Galimberti (2013).

2.2 A probabilistic Bayesian learning framework

Consider a framework where agents are interested in estimating the value of an unobservable variable θ_t, $t \geq 1$. Nature draws the value θ_t at some time $t = 0$ from an improper uniform distribution over $\mathbb{R}$. Consistently, agents have a flat (uninformative) prior on its value at time $t = 1$. Nature can also re-draw, with some fixed and known probability $0 \leq \pi \leq 1$, a new value for the variable, again from an improper distribution over $\mathbb{R}$, at the beginning of each period $t > 1$. At every period $t \geq 1$ agents receive a signal x_t on the value of θ_t, with the form

$$x_t = \theta_t + v_t, \tag{5}$$

where v_t is an i.i.d. random variable, normally distributed with zero mean and constant variance σ_v^2.

I first define

$$\bar{x}_{j\leq t}^t = \frac{1}{t-j+1}\sum_{z=j}^{t} x_z,$$

the best estimate of θ_t at time t if Nature had last re-drawn at (the beginning of) time $j \leq t$ and agents knew it. This is simply the mean of the sample of relevant observations for the signal, since the last change in θ_t took place.

Given that agents don't know if and when a change in the estimated variable took place, their posterior mean for it at time t is then given by

$$\tilde{x}_t = \sum_{j=1}^{t} a_j^t \bar{x}_{j\leq t}^t$$

where

$$a_1^t = (1-\pi)^{t-1} \tag{6}$$
$$a_j^t = (1-\pi)^{t-j}\,\pi, t \geq j > 1, \tag{7}$$

with

$$\sum_{j=1}^{t} a_j^t = 1.$$

The coefficients a_j^t capture the probability that each (truncated) series $\bar{x}_{j\leq t}^t$ is the appropriate one for computing the conditional expected value of θ_t, that is, the probability that Nature re-drew at the beginning of time j and never after. Clearly, (6)-(7) are the same as (3)-(4) for $\pi = g$.

It is then possible to rewrite the posterior $\tilde{x}_t$ as a weighted sum of current and past values of x_t as

$$\tilde{x}_t = \sum_{j=1}^{t} h_j^t x_j, \tag{8}$$

where

$$h_1^t = \frac{(1-\pi)^{t-1}}{t} \tag{9}$$
$$h_j^t = h_{j-1}^t + \frac{(1-\pi)^{t-j}\,\pi}{t-j+1}, t \geq j > 1. \tag{10}$$

It can be shown that $\sum_{j=1}^{t} h_j^t = 1$. Clearly if $\pi = 1$ (θ_t changes for sure every period), $h_j^t = 0$ for $j < t$ and $h_j^t = 1$ for $j = t$: only the last observation matters. If instead $\pi = 0$ (θ_t constant) then all observations receive the same weight $1/t$. This gives rise to a decreasing gain algorithm, implementing recursive least squares (equivalent to stochastic gradient in this case)

$$\tilde{x}_t = \tilde{x}_{t-1} + \frac{1}{t}\left(x_t - \tilde{x}_{t-1}\right)$$

with $\tilde{x}_0 = 0$ (that is, $\tilde{x}_1 = x_1$), or, in non-recursive form,

$$\tilde{x}_t = \frac{1}{t}\sum_{z=1}^{t} x_z, \tag{11}$$

which is simply the sample mean.

To better understand the weighting structure defined by (9)-(10), I propose Fig (1).

Observation x_1 is relevant for inference about the current value of θ_t only if Nature never re-

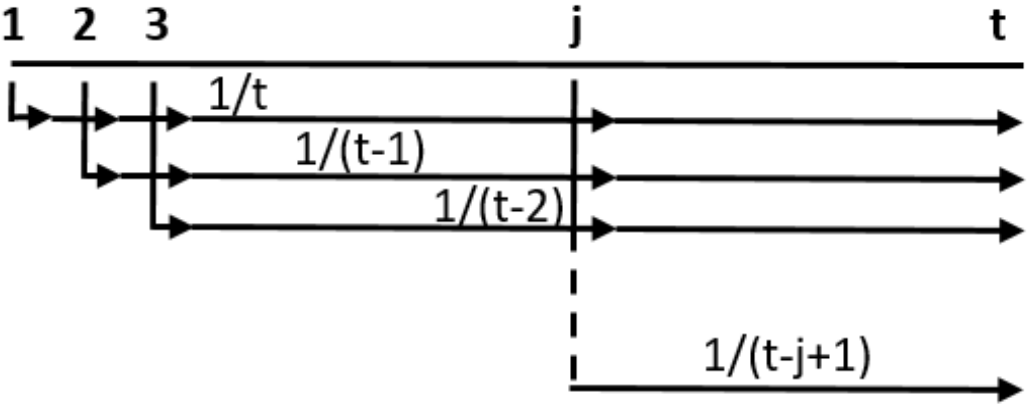

Figure 1: Weighting structure of signals.

drew over the whole sample period from 1 to t, which happened with probability $(1-\pi)^{t-1}$: in such case each observation in that sample should be weighted equally, with weight $1/t$. Observation x_2 is relevant if Nature never re-drew (again, with weight $1/t$), which happened with probability $(1-\pi)^{t-1}$, or if it re-drew at the beginning of period 2 and never after (and in this case, with weight $\frac{1}{t-1}$), which happened with probability $(1-\pi)^{t-2}\pi$. And so on.

3 A comparison

In light of the proposed framework, it is instructive to analyze the relationship between the adaptive learning gain g and the probability π in the Bayesian learning model. The gain parameter in an adaptive learning algorithm determines the weight put on past observations: with a decreasing gain $1/t$, all observations receive equal weight; with a constant gain g, instead, the weight decays exponentially with past observations. A similar interpretation can be given to π, which represents the probability of a change in the variable θ_t happening at each time t: this determines the probability that each observation from time j, $0 < j \leq t$ is relevant for time t inference, which, together with the number of observations, determines individual weights.

The weighting structure represented by (9)-(10) cannot be generated by a CG algorithm for finite t. Nevertheless, it provides a means to interpret the weighting implied by such algorithm. Clearly, if one sets $\pi = g$ then $b_j^t = a_j^t$: if the constant gain is to be interpreted as

the probability π, the weight put on individual observations through the CG algorithm are the weights put on past truncated series of observations in the probabilistic Bayesian setting. In such setting, weights on individual observations are instead given by (9)-(10), which, in a non-recursive way, can be rewritten as

$$\begin{aligned} h_1^t &= \frac{(1-\pi)^{t-1}}{t} \\ h_j^t &= \frac{(1-\pi)^{t-1}}{t} + \sum_{m=2}^{j} \frac{(1-\pi)^{t-m}\,\pi}{t-m+1}. \end{aligned}$$

While the weighting structure on individual observations in the Bayesian framework is more convoluted than that in the CG algorithm, both b_j^t and the leading term in h_j^t (represented by $\frac{(1-\pi)^{t-j}\pi}{t-j+1}$) decay exponentially, leading to similar weight profiles on older observations. In fact, for $\pi = g$, the leading term in h_j^t is equal to $\frac{b_j^t}{t-j+1}$.

Figure 2 shows b_j^t and h_j^t, computed for $\pi = g = 0.025$ with $t = 100$. Figure 3 then shows the same series, but for $t = 1,000$. It can be seen that as t increases, b_j^t and h_j^t get closer to each other for small values of j, while for high values of j (that is, for observations closer to the time of estimation) the difference between the two terms remains largely the same. Weights b_j^t, $j \geq 1$, are independent of t (they depend instead on $t-j$; that is, $b_j^t = b_{j+k}^{t+k}$). The same is not exactly true for h_j^t, though quantitatively it is indeed the case that $h_j^t \simeq h_{j+k}^{t+k}$. This is due to the fact that the leading $j-1$ terms out of the total $j+k$ terms in h_{j+k}^{t+k} are the same as the leading $j-1$ terms out of the total j terms in h_j^t (the j^{th} term differs by π), with the additional k terms in h_{j+k}^{t+k} negligible in size. Thus the final end of the b and h curves tend to remain at the same distance as t increases. The two curves, instead, get closer and closer to each other on their initial part as t increases, because both tend to zero (weights on observations farther back in time converge to zero under both weighting structures).

It can be seen that, despite being derived in different frameworks, the shape of the two weighting structures is remarkably similar, leading to similar weighting on past information in the two cases. Figure 4 shows the difference $\Delta^t = b^t - h^t$ (where b^t and h^t are the vectors $\left\{b_j^t\right\}_{j=1}^t$ and $\left\{h_j^t\right\}_{j=1}^t$) for $t = 1,000$ and $\pi = g = 0.025$.

4 Constant gains in the empirical literature

Using the framework developed above, one can interpret the constant gain coefficients that have been found to fit the data well in empirical macroeconomic studies in terms of the

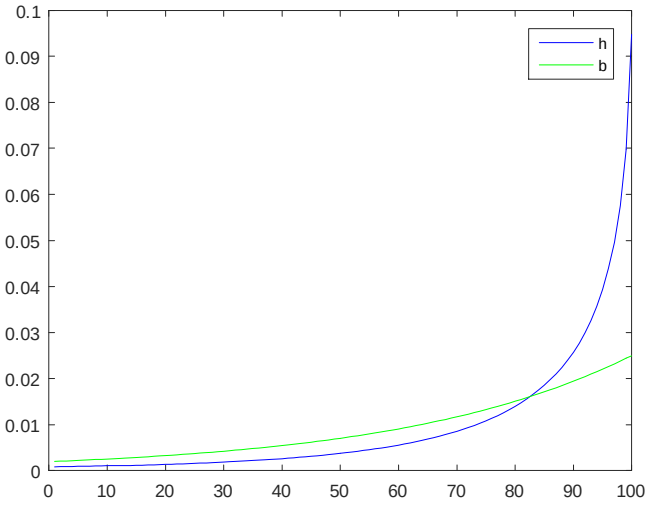

Figure 2: Values of b_j^t and h_j^t for $t = 100$.

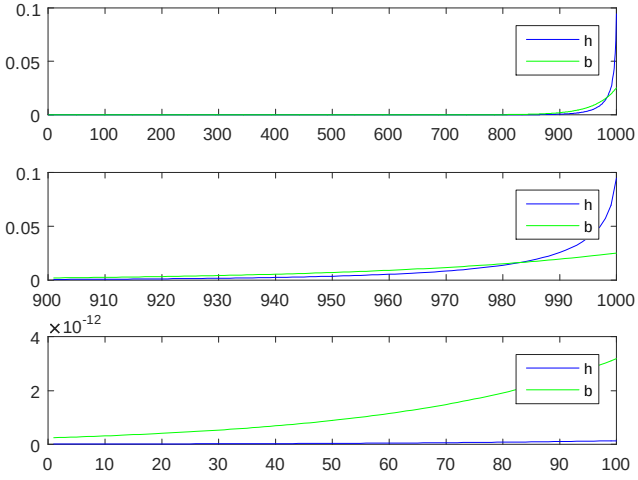

Figure 3: Values of b_j^t and h_j^t for $t = 1000$.

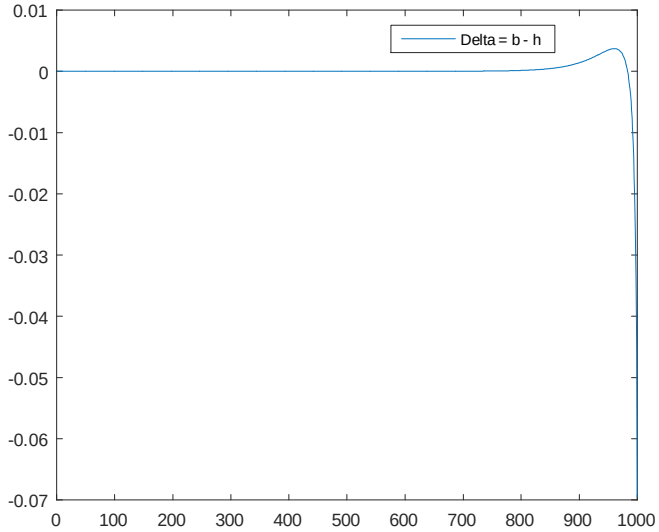

Figure 4: Values of Δ_j^{1000} for $g = 0.025$.

implied probability of changes in the estimated parameters. Typical values used (estimated or calibrated) in the empirical literature for the constant gain range from close to zero to over 0.2, though most studies use values between 0.01 and 0.1, as reported in Berardi and Galimberti (2017).

One can compute the implied probability of changes in the estimated parameters that corresponds to a specific gain coefficient by finding the π that, for a given g, minimizes the sum of (squared) deviations between the weighting structure implied by the gain and the weighting structure of the Bayesian framework. That is, one can compute

$$\hat{\pi}^t(g) = \arg\min_{\pi} \Delta^t(g)'\Delta^t(g)$$

where the notation for $\hat{\pi}^t(g)$ and $\Delta^t(g)$ makes explicit the dependence on both t and g. Fixing g, one can find the implied probability for a certain gain coefficient as a function of the number of observations. Figure 5 shows such measure for $g = 0.025$. It can be seen that for large enough values of t, $\hat{\pi}^t(g = 0.025)$ stabilizes and becomes constant. One can thus compute the value of $\hat{\pi}^t(g)$ for large t,[1] obtaining a function that gives the implied (asymptotic) probability $\hat{\pi}$ for any value of g. In particular, I restrict the range of g between

[1] I set $t = 1,000$ in the computations.

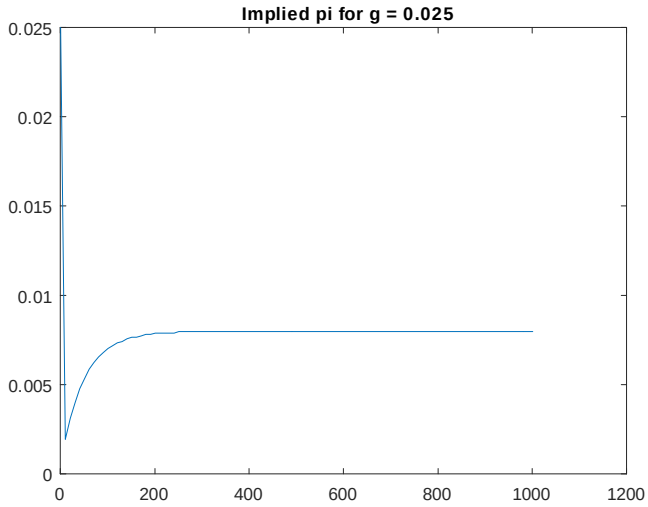

Figure 5: Values of $\hat{\pi}^t(g = 0.025)$.

0.01 and 0.1, which contains most values used in the empirical literature. Figure 6 shows the results. It can be seen that, for gains between 0.01 and 0.1, the implied probability of changes in the estimated parameter(s) each period ranges from 0.31% ($g = 0.01$) to 3.59% ($g = 0.1$).

5 Conclusions

This paper has proposed a probabilistic Bayesian framework which allows for the interpretation of the weighting structure of past observations implied by the CG learning algorithm in terms of the probability of changes in the estimated parameters. It is then possible to map the gain coefficients used in the empirical literature into implied probabilities. For example, a gain of 0.025 corresponds to a probability of changes in the estimated parameter of 0.31% every period, while a gain of 0.1 corresponds to a probability of 3.59%.

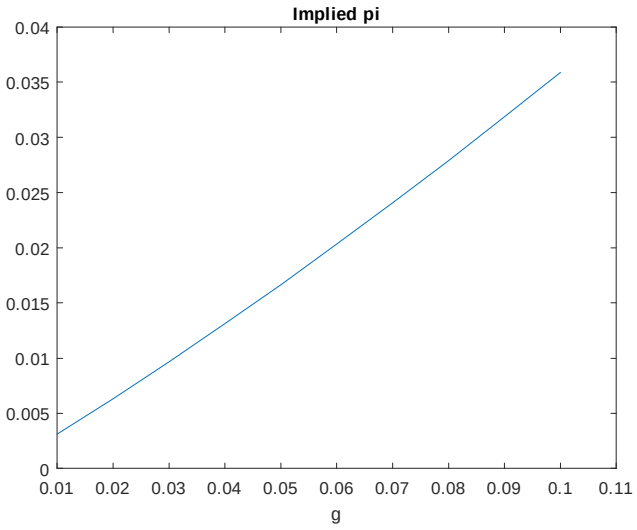

Figure 6: Values of $\hat{\pi}^{1000}(g)$

References

[1] Berardi, M., Galimberti, J.K., 2013. A note on exact correspondences between adaptive learning algorithms and the kalman filter. Economics Letters 118, 139–142.

[2] Berardi, M., Galimberti, J.K., 2017. Empirical calibration of adaptive learning. Journal of Economic Behavior and Organization 144, 219-237.

[3] Evans, G.W., Honkapohja, S., 2001. Learning and Expectations in Macroeconomics. Princeton University Press.

[4] Eusepi, S., Preston, B., 2011. Expectations, learning, and business cycle fluctuations. American Economic Review 101, 2844–2872.

[5] Milani, F., 2011. Expectation shocks and learning as drivers of the business cycle. The Economic Journal 121, 379–401.

[6] Muth, J.F., 1960. Optimal Properties of Exponentially Weighted Forecasts. Journal of the American Statistical Association 55, 299-306.

[7] Sargent, T.J., 1999. The Conquest of American Inflation. Princeton, NJ: Princeton University Press.

[8] Sargent, T.J., Williams, N., Zha, T., 2006. Shocks and government beliefs: the rise and fall of American inflation. American Economic Review 96, 1193-1224.